JPO: Juris Policy Optimization for Structured Legal Reasoning in Criminal Judgment Prediction

Anonymous ACL submission

Abstract

Criminal judgment prediction requires models to infer statutory articles, charges, and sentencing outcomes from case facts. Unlike standard classification tasks, it involves a structured reasoning process in which statutes should support facts, charges should justify statutes, and sentencing outcomes should remain consistent with charges. Existing approaches mainly optimize final labels, with limited supervision or optimization for intermediate legal reasoning. We propose **Juris Policy Optimization (JPO)**, a post-training framework for structured legal reasoning in Chinese criminal judgment prediction. JPO first uses teacher-generated rationales to supervise a standardized four-step reasoning process, and then applies reinforcement learning with a composite reward over legal prediction quality, reasoning structure completeness, and cross-step consistency. JPO further introduces token-level advantage reweighting and adaptive clipping for legally salient reasoning segments. Experiments on multiple open-source language models and three Chinese legal benchmarks show that JPO consistently improves both judgment prediction and reasoning quality over supervised fine-tuning and reinforcement learning baselines.

1 Introduction

Legal judgment prediction (LJP) aims to infer judicial outcomes from case facts, including applicable statutory articles, criminal charges, and sentencing decisions. As a core task in legal NLP, it has been studied to support legal information processing, case analysis, and intelligent legal assistance (Xiao et al., 2018; Chalkidis et al., 2022; Guha et al., 2023). In criminal cases, however, judgment prediction is not merely a flat mapping from facts to labels. A plausible decision must remain coherent across several reasoning stages: the cited statutory provisions should support legally relevant facts, the predicted charge should justify statutory provisions, and the final sentence should be compatible with that charge.

This structured nature is particularly important in Chinese criminal law, where judicial decisions are typically organized around factual findings, statutory articles, charges, and sentencing outcomes. As a result, high final-label accuracy alone is insufficient: a model may predict the correct charge while relying on unsupported statutes, or output a plausible sentence without a coherent legal basis.

Recent large language models (LLMs) have shown strong performance on reasoning-intensive tasks and motivated growing interest in legal-domain applications (OpenAI et al., 2024; Touvron et al., 2023; Bai et al., 2023; Wei et al., 2022; Kojima et al., 2023; Guha et al., 2023). In legal judgment prediction, LLMs provide a natural way to generate both predictions and rationales. However, adapting general-purpose LLMs to criminal judgment prediction remains difficult for two reasons. First, most existing LJP datasets provide case facts and final labels, but little supervision for intermediate reasoning (Xiao et al., 2018; Fei et al., 2024). Models trained on such data can learn shallow correlations between facts and outcomes rather than legally grounded reasoning patterns (Liu et al., 2024; Shi et al., 2025). Second, existing post-training methods mainly optimize final outputs or general preference signals (Ouyang et al., 2022; Rafailov et al., 2023; Dai et al., 2025; Cai et al., 2025), but do not explicitly model the structured dependencies that define legal adjudication.

To address this gap, we propose **Juris Policy Optimization (JPO)**, a post-training framework for structured legal reasoning in Chinese criminal judgment prediction. JPO follows a two-stage design. In the first stage, a stronger teacher model generates rationales following a standardized four-step structure: fact extraction, statutory analysis, charge determination, and sentence prediction.

These rationales are used to supervise structured reasoning generation during fine-tuning. In the second stage, we apply reinforcement learning with a composite reward that jointly captures legal prediction accuracy, reasoning structure completeness, and cross-step consistency across facts, statutes, charges, and sentences. To provide finer-grained optimization signals, JPO further introduces token-level advantage reweighting and adaptive clipping for legally salient reasoning segments.

We evaluate JPO on multiple open-source LLM backbones and three Chinese legal benchmarks. Experimental results show that JPO consistently improves both judgment prediction performance and reasoning quality over supervised fine-tuning and standard reinforcement learning baselines. The gains are especially clear on compact models, suggesting that structured legal-reasoning alignment can substantially strengthen smaller open-source models for criminal judgment prediction.

Our contributions are summarized as follows:

- We propose **Juris Policy Optimization (JPO)**, a post-training framework for structured legal reasoning in criminal judgment prediction.
- We introduce a composite reinforcement learning objective that jointly optimizes legal prediction quality, reasoning completeness, and cross-step consistency.
- We develop token-level advantage reweighting and adaptive clipping for legally salient reasoning segments.
- Experiments on multiple open-source backbones and three Chinese legal benchmarks show that JPO consistently improves both judgment prediction and reasoning quality over strong baselines.

2 Related Work

Legal judgment prediction (LJP) is a central task in legal NLP, aiming to infer judicial outcomes such as applicable statutes, charges, and sentencing results from case facts. Early work on Chinese criminal judgment prediction was largely benchmark-driven, with datasets such as CAIL establishing standard settings for article prediction, charge prediction, and prison term prediction (Xiao et al., 2018). More recent studies have argued that accurate legal prediction requires more than final-label matching and should better capture the reasoning process connecting facts, legal rules, and judicial outcomes (Lee et al., 2025; Liu et al., 2024; Shi et al., 2025). Still, most existing approaches primarily rely on supervision over final labels, without explicitly optimizing a standardized multi-stage reasoning process.

Large language models have shown strong performance on reasoning-intensive tasks and have motivated increasing interest in legal-domain adaptation and evaluation (OpenAI et al., 2024; Wei et al., 2022; Guha et al., 2023; Fei et al., 2024). In legal judgment prediction, LLMs provide a natural way to jointly generate predictions and rationales. However, adapting general-purpose LLMs to criminal judgment prediction remains difficult because most legal datasets provide limited supervision for intermediate reasoning, and unconstrained rationale generation does not by itself ensure legally grounded reasoning or cross-step coherence.

In parallel, post-training methods such as RLHF and DPO have become standard approaches for aligning language models with task objectives and preference signals (Ouyang et al., 2022; Rafailov et al., 2023). Recent work has begun to extend these ideas to legal-domain reasoning through domain-specific post-training and reinforcement learning (Dai et al., 2025; Cai et al., 2025). Our work is most closely related to this line of research, but differs in a key respect: instead of optimizing only final answers or general legal preferences, we explicitly optimize the structured dependency chain from facts to statutes, charges, and sentencing outcomes in Chinese criminal judgment prediction.

3 Method

We propose **Juris Policy Optimization (JPO)**, a two-stage post-training framework for structured legal reasoning in Chinese criminal judgment prediction. The central idea is to align a language model not only with the final judicial outcome, but also with the intermediate reasoning structure connecting facts, statutory articles, charges, and sentencing decisions. As illustrated in Figure 1, JPO first introduces structured supervision through teacher-generated four-step legal rationales, and then further improves the model with reinforcement learning under a composite reward that captures legal correctness, reasoning completeness, and cross-step consistency. JPO further incorporates token-level advantage reweighting and adaptive policy clipping to better allocate optimization

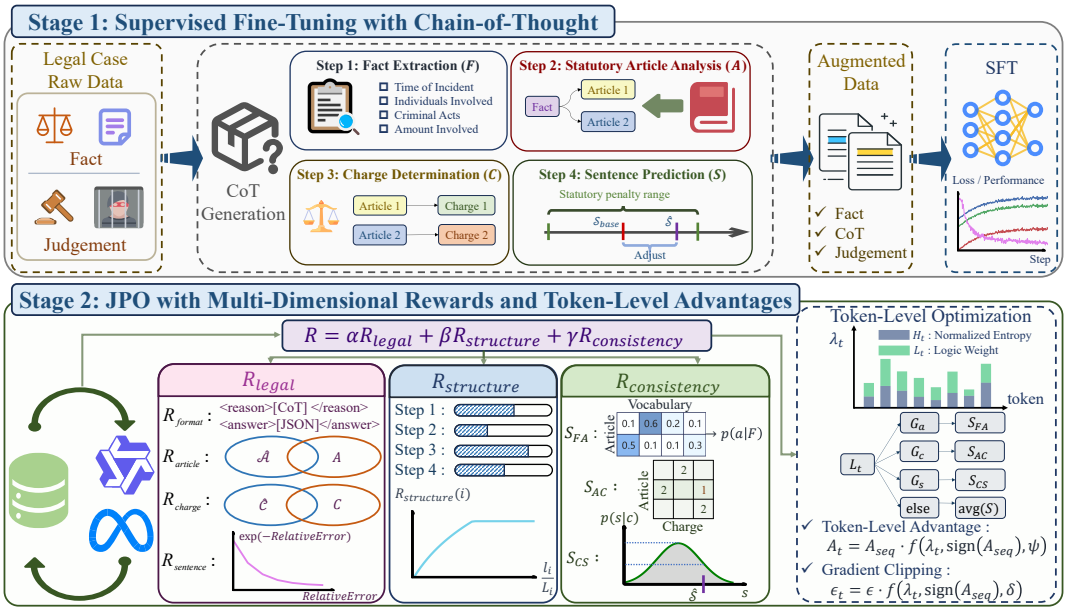

Figure 1: **Overview of JPO.** Structured SFT with teacher-generated four-step rationales, followed by reinforcement learning with rewards for legal accuracy, reasoning completeness, and cross-step consistency.

signals within long legal responses.

3.1 Task Definition

Given a criminal case described by a fact sequence

$$F = \{f_1, f_2, \ldots, f_n\}, \tag{1}$$

the goal of legal judgment prediction is to infer the judicial outcome

$$Y = (A, C, S), \tag{2}$$

where A denotes the applicable statutory articles, C denotes the criminal charge, and S denotes the sentencing outcome.

Beyond the final outcome, JPO explicitly models the reasoning trace as

$$Z = (G_f, G_a, G_c, G_s), \tag{3}$$

where G_f, G_a, G_c, and G_s correspond to fact extraction, statutory analysis, charge determination, and sentence prediction, respectively. These four segments define an observable reasoning template aligned with the adjudication trajectory

$$F \rightarrow A \rightarrow C \rightarrow S. \tag{4}$$

Accordingly, JPO treats criminal judgment prediction as a structured generation problem in which the reasoning process Z and the final outcome Y are learned jointly.

3.2 Stage I: Structured Supervised Fine-Tuning

Most legal judgment prediction datasets provide case facts and final labels but do not annotate intermediate legal reasoning. JPO addresses this limitation by introducing structured reasoning supervision in the first stage.

For each training example, a stronger instruction-following teacher model generates a reasoning trace Z following the four-stage structure (G_f, G_a, G_c, G_s). The student model is then trained to generate both the structured trace and the final prediction conditioned on F:

$$\mathcal{L}_{\text{SFT}} = -\log P_\theta(Z, Y \mid F), \tag{5}$$

where θ denotes the model parameters.

This stage teaches the model a stable and interpretable reasoning template, and provides initialization for the subsequent reinforcement learning stage.

3.3 Stage II: Juris Policy Optimization

Although structured SFT teaches the model the desired reasoning format, it does not guarantee legally grounded reasoning. We therefore introduce a second stage based on reinforcement learning, whose reward explicitly captures both correct outcomes and coherent reasoning:

$$\mathcal{R} = \alpha R_{\text{legal}} + \beta R_{\text{structure}} + \gamma R_{\text{consistency}}, \tag{6}$$

where the three components measure legal prediction quality, structural completeness, and cross-step consistency, respectively.

3.4 Legal Prediction Reward

The legal prediction reward evaluates whether the generated output matches the target judicial outcome. Since criminal judgment prediction contains multiple decision targets, we decompose this reward into four terms:

$$R_{\text{legal}} = \alpha_1 R_{\text{format}} + \alpha_2 R_{\text{article}} + \alpha_3 R_{\text{charge}} + \alpha_4 R_{\text{sentence}}. \tag{7}$$

The format reward R_{format} checks whether the generated response follows the required response schema, including both the reasoning trace and the final structured answer.

For article prediction, let $\hat{A}$ denote the predicted set of statutory articles. We define

$$R_{\text{article}} = \frac{|\hat{A} \cap A|}{|\hat{A}|} \cdot \frac{|\hat{A} \cap A|}{|A|}. \tag{8}$$

This reward is high only when the predicted article set is both precise and complete with respect to the gold set. In practice, when $\hat{A}$ is empty, we define $R_{\text{article}} = 0$.

Similarly, let $\hat{C}$ denote the predicted set of charges. The charge reward is

$$R_{\text{charge}} = \frac{|\hat{C} \cap C|}{|\hat{C}|} \cdot \frac{|\hat{C} \cap C|}{|C|}. \tag{9}$$

This formulation again encourages predictions that are simultaneously precise and complete. Likewise, when $\hat{C}$ is empty, we define $R_{\text{charge}} = 0$.

For sentencing prediction, exact matching is often too strict because sentencing is continuous and sensitive to case-specific aggravating and mitigating factors. We therefore define a smooth reward based on relative error:

$$R_{\text{sentence}} = \exp\left(-\xi \frac{|\hat{S} - S|}{S}\right), \tag{10}$$

where $\hat{S}$ denotes the predicted sentence. This formulation preserves graded supervision even when the predicted sentence is only partially accurate.

3.5 Reasoning Structure Reward

Correct final predictions alone are insufficient if the generated reasoning is incomplete or degenerate. We therefore introduce an explicit reward for outputs that preserve the expected four-stage organization:

$$R_{\text{structure}} = \sum_{k \in \{f,a,c,s\}} w_k \, \mathbb{I}(G_k) \Big[\min(\log(1 + \eta \ell_k), 1) \Big]. \tag{11}$$

where $\mathbb{I}(G_k)$ indicates whether the corresponding reasoning segment appears in the generated trace.

The normalized segment length is defined as

$$\ell_k = \frac{l_k}{L_k}, \tag{12}$$

where l_k is the actual length of the generated segment and L_k is the target minimum length for that segment.

This reward plays two complementary roles: the indicator term encourages the model to include all required reasoning stages, while the length term discourages empty or trivial segment generation.

3.6 Logical Consistency Reward

A response may be structurally complete yet still remain legally incoherent across reasoning steps. We therefore define a consistency reward over the three local transitions in the legal reasoning chain:

$$R_{\text{consistency}} = \frac{1}{3}\left(S_{FA} + S_{AC} + S_{CS}\right). \tag{13}$$

Here, S_{FA}, S_{AC}, and S_{CS} denote fact-to-article, article-to-charge, and charge-to-sentence consistency, respectively. These terms are designed as lightweight and computable proxies for the three adjacent reasoning transitions.

Fact-to-article consistency. We estimate whether the predicted statutory articles are supported by the case facts using a Naive Bayes article predictor trained on the corpus. Let $\mathcal{A}$ denote the full article space, and let $P(a \mid F)$ denote the posterior probability assigned to article $a \in \mathcal{A}$ for case facts F:

$$P(a \mid F) = \frac{P(a) \prod_{f \in F} P(f \mid a)}{\sum_{a' \in \mathcal{A}} P(a') \prod_{f \in F} P(f \mid a')}. \tag{14}$$

Given the predicted article set $\hat{A}$, we define the fact-to-article consistency score as

$$S_{FA} = \frac{1}{|\hat{A}|} \sum_{a \in \hat{A}} P(a \mid F). \tag{15}$$

When $\hat{A}$ is empty, we define $S_{FA} = 0$.

This score is high when the predicted articles are well supported by the observed case facts.

Article-to-charge consistency. We next evaluate whether the predicted charge is compatible with the predicted statutory articles through an article–charge association matrix

$$D \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{C}|}, \tag{16}$$

where $\mathcal{A}$ and $\mathcal{C}$ denote the full article space and charge space, respectively, and $D_{a,c}$ measures the strength of association between article a and charge c.

In our implementation, $D_{a,c}$ is defined by a lightweight rule-based mapping derived from statutory references and charge annotations in the training data: $D_{a,c} = 2$ if article a directly governs charge c, $D_{a,c} = 1$ if the relation is relevant but indirect, and $D_{a,c} = 0$ otherwise.

Given the predicted article set $\hat{A}$ and predicted charge set $\hat{C}$, the corresponding consistency score is

$$S_{AC} = \frac{\sum_{a\in\hat{A}} \max_{c\in\hat{C}} D_{a,c}}{|\hat{A}| \max(D)}. \tag{17}$$

When $\hat{A}$ or $\hat{C}$ is empty, we define $S_{AC} = 0$.

This term favors reasoning traces in which the predicted charge is supported by the cited statutory basis.

Charge-to-sentence consistency. Finally, we evaluate whether the predicted sentence is plausible under the predicted charge. For each charge c, we estimate a charge-conditioned sentence distribution from the training data using a truncated normal distribution:

$$p(s \mid c) = \frac{\phi\left(\frac{s-\mu_c}{\sigma_c}\right)}{\sigma_c\left[\Phi\left(\frac{b_c-\mu_c}{\sigma_c}\right) - \Phi\left(\frac{a_c-\mu_c}{\sigma_c}\right)\right]}, \tag{18}$$

where $[a_c, b_c]$ is the observed valid sentencing range for charge c in the training corpus, and μ_c and σ_c are the empirical mean and standard deviation of the sentence distribution for that charge.

The resulting consistency score is

$$S_{CS} = \frac{1}{|\hat{C}|} \sum_{c\in\hat{C}} \frac{p(\hat{S} \mid c)}{\max_s p(s \mid c)}. \tag{19}$$

When $\hat{C}$ is empty, we define $S_{CS} = 0$.

This term measures whether the predicted sentence lies in a plausible region of the charge-conditioned distribution.

3.7 Token-Level Advantage Reweighting

The rewards above are computed at the sequence level, but different tokens contribute unequally to legal reasoning quality. We therefore introduce token-level advantage reweighting. Let A_{seq} denote the sequence-level advantage. For token t,

$$A_t = A_{\text{seq}} \cdot f(\lambda_t, \text{sign}(A_{\text{seq}}), \psi), \tag{20}$$

where $\lambda_t \in [0, 1]$ denotes the importance weight of token t, and ψ controls the reweighting strength.

The importance score combines two complementary signals:

$$\lambda_t = \zeta H_t + (1 - \zeta) L_t, \tag{21}$$

where H_t denotes the normalized entropy weight and L_t denotes the legal logic weight.

Entropy weight. The entropy term measures model uncertainty at token t. Let $p_t(v)$ denote the predicted probability of vocabulary item v at decoding step t. The token entropy is

$$\mathcal{H}_t = -\sum_{v\in\mathcal{V}} p_t(v) \log p_t(v), \tag{22}$$

where $\mathcal{V}$ is the vocabulary.

Since entropy depends on vocabulary size and decoding dynamics, we normalize it as

$$H_t = \frac{\mathcal{H}_t}{\max_{\tau=1}^{T} \mathcal{H}_\tau}. \tag{23}$$

A larger H_t indicates higher uncertainty and thus a potentially stronger need for learning signal.

Legal logic weight. The legal logic weight depends on the reasoning segment to which token t belongs. Tokens are assigned to segments according to their position in the generated reasoning trace $Z = (G_f, G_a, G_c, G_s)$. We define

$$L_t = \begin{cases} S_{FA}, & t \in G_a \\ S_{AC}, & t \in G_c \\ S_{CS}, & t \in G_s \\ \frac{1}{3}(S_{FA} + S_{AC} + S_{CS}), & \text{otherwise.} \end{cases} \tag{24}$$

Given λ_t, the reweighting function is

$$f(\lambda_t, \text{sign}(A_{\text{seq}}), \psi) = \begin{cases} 1 + \psi(\lambda_t - \bar{\lambda}), & A_{\text{seq}} > 0 \\ 1 - \psi(\lambda_t - \bar{\lambda}), & A_{\text{seq}} < 0 \\ 1, & A_{\text{seq}} = 0 \end{cases} \tag{25}$$

where

$$\bar{\lambda} = \frac{1}{T}\sum_{\tau=1}^{T}\lambda_\tau \tag{26}$$

is the average token importance of the sequence and T is the sequence length.

When A_{seq} is positive, above-average important tokens receive larger positive updates. When A_{seq} is negative, the same tokens are penalized less aggressively, which helps preserve legally salient content during policy optimization.

3.8 Adaptive Policy Clipping

Standard clipped policy optimization uses a uniform clipping threshold for all tokens. We instead introduce token-aware clipping:

$$\epsilon_t = \epsilon \cdot f(\lambda_t, \text{sign}(A_{\text{seq}}), \delta), \tag{27}$$

where ϵ is the base clipping coefficient and δ controls the adaptation strength. This design allows larger updates on legally important tokens while preserving more conservative optimization on less informative parts of the output.

3.9 Policy Optimization Objective

Given the token-level advantages A_t and token-dependent clipping coefficients ϵ_t, we optimize the policy using a clipped policy gradient objective:

$$\mathcal{L}_{\text{RL}} = \mathbb{E}_t\Big[\min\big(r_t A_t,\ \text{clip}(r_t, 1-\epsilon_t, 1+\epsilon_t)A_t\big) - \beta\mathcal{D}_{KL}\big[\pi_\theta(a_t \mid s_t)\|\pi_{\theta_{\text{ref}}}(a_t \mid s_t)\big]\Big]. \tag{28}$$

Here, r_t denotes the probability ratio between the current policy and the old policy.

The objective combines token-level reweighted advantages, token-aware clipping, and KL regularization with respect to a reference policy. The KL term prevents the policy from drifting excessively away from the supervised initialization and helps preserve stable legal-domain generation during post-training.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate JPO on three Chinese criminal judgment prediction benchmarks: JPO-Dataset, CAIL2018 (Xiao et al., 2018), and LawBench (Fei et al., 2024). JPO-Dataset is a newly constructed dataset derived from public judicial documents collected from China Judgments Online. We introduce it because widely used Chinese criminal judgment prediction benchmarks were built from relatively earlier judicial documents and may not fully reflect more recent criminal case distributions and legal expressions. Following the general construction protocol of CAIL2018, we re-collect recent criminal cases and extract fact descriptions, applicable statutory articles, charges, and sentencing outcomes, together with normalization of article references, charge labels, and sentencing expressions. JPO-Dataset is used for structured supervised fine-tuning and reinforcement learning, while CAIL2018 and LawBench are used as external evaluation benchmarks. Basic statistics are shown in Table 2.

Models. We conduct experiments on five open-source backbones: Qwen2.5-3B/7B-Instruct, Qwen3-4B-Instruct, Llama-3.2-3B-Instruct, and Llama-3-8B-Instruct (Qwen et al., 2025; Yang et al., 2025; Grattafiori et al., 2024). For the 4B/7B/8B models, we report Pre-trained, SFT, and JPO. For the two representative 3B backbones, we further include RL-only, Vanilla PPO (Schulman et al., 2017), LegalΔ (Dai et al., 2025), and Issue Tree Rubrics (Cai et al., 2025), together with component ablations. We also report several proprietary models as zero-shot reference systems, including DeepSeek-V3.2, Qwen3-32B, GPT-5.2, and Claude-Sonnet-4.5. These systems are included only as reference points for model capability and are not directly comparable to our trained open-source systems because they are evaluated in a zero-shot setting without task-specific post-training on JPO-Dataset.

Evaluation Metrics. We report five metrics in Table 1. Article prediction and charge prediction are evaluated using macro-F1. Sentence prediction is evaluated by a relative-error-based score:

$$\text{Score}_{\text{sentence}} = \exp\left(-\xi \cdot \frac{|\hat{S} - S|}{S}\right), \tag{29}$$

where $\xi = 3$ during evaluation, $\hat{S}$ denotes the predicted sentence, and S denotes the gold sentence. We additionally report *4-Step Completeness*, which measures whether the generated response explicitly contains all four reasoning stages required by our framework, and *Full-Chain Consistency*, which measures whether the generated reasoning remains coherent across fact extraction, statutory analysis, charge determination, and sentence prediction. The

Method	Set.	JPO-Dataset					CAIL2018					LawBench				
		Art. (F1)	Charge (F1)	Sent. (Score)	4-Step (Comp.)	Full (Chain)	Art. (F1)	Charge (F1)	Sent. (Score)	4-Step (Comp.)	Full (Chain)	Art. (F1)	Charge (F1)	Sent. (Score)	4-Step (Comp.)	Full (Chain)
Proprietary Models (Zero-Shot Reference)																
DeepSeek-V3.2	Zero-Shot	0.924	0.908	0.512	0.958	0.771	0.896	0.875	0.481	0.942	0.738	0.864	0.841	0.439	0.925	0.706
Qwen3-32B	Zero-Shot	0.917	0.911	0.504	0.961	0.759	0.887	0.882	0.473	0.945	0.722	0.858	0.835	0.428	0.931	0.694
GPT-5.2	Zero-Shot	0.912	0.894	0.489	0.947	0.744	0.879	0.861	0.455	0.933	0.705	0.845	0.829	0.412	0.916	0.683
Claude-Sonnet-4.5	Zero-Shot	0.908	0.891	0.493	0.952	0.751	0.882	0.866	0.462	0.937	0.713	0.849	0.824	0.417	0.918	0.687
Open-Source Results																
Qwen3-4B-Instruct	Pre-trained	0.521	0.468	0.174	0.315	0.208	0.485	0.442	0.158	0.291	0.187	0.463	0.419	0.142	0.264	0.175
	SFT	0.884	0.858	0.405	0.902	0.652	0.856	0.831	0.381	0.882	0.627	0.834	0.809	0.364	0.857	0.598
	JPO	**0.931**	**0.916**	**0.542**	**0.966**	**0.789**	**0.903**	**0.888**	**0.514**	**0.951**	**0.758**	**0.872**	**0.861**	**0.485**	**0.935**	**0.724**
Qwen2.5-7B-Instruct	Pre-trained	0.618	0.573	0.211	0.442	0.267	0.584	0.539	0.194	0.408	0.241	0.551	0.508	0.182	0.377	0.226
	SFT	0.893	0.871	0.422	0.918	0.681	0.865	0.842	0.399	0.895	0.655	0.841	0.824	0.377	0.875	0.628
	JPO	**0.937**	**0.928**	**0.551**	**0.973**	**0.806**	**0.908**	**0.897**	**0.521**	**0.958**	**0.774**	**0.879**	**0.868**	**0.493**	**0.942**	**0.741**
Llama-3-8B-Instruct	Pre-trained	0.594	0.552	0.203	0.418	0.252	0.561	0.524	0.187	0.389	0.229	0.537	0.495	0.175	0.362	0.214
	SFT	0.871	0.847	0.396	0.897	0.641	0.846	0.822	0.375	0.872	0.614	0.823	0.798	0.354	0.851	0.585
	JPO	**0.926**	**0.912**	**0.534**	**0.962**	**0.781**	**0.897**	**0.884**	**0.505**	**0.946**	**0.749**	**0.868**	**0.856**	**0.477**	**0.928**	**0.716**
Detailed 3B Analysis																
Qwen2.5-3B-Instruct	Pre-trained	0.451	0.382	0.114	0.284	0.198	0.426	0.359	0.103	0.258	0.177	0.402	0.334	0.092	0.231	0.165
	RL-only	0.782	0.744	0.295	0.671	0.482	0.755	0.718	0.273	0.644	0.457	0.732	0.695	0.258	0.615	0.434
	SFT	0.873	0.847	0.391	0.895	0.622	0.848	0.823	0.372	0.874	0.596	0.825	0.801	0.349	0.848	0.568
	Vanilla PPO	0.896	0.869	0.454	0.911	0.678	0.871	0.845	0.428	0.891	0.649	0.846	0.822	0.405	0.867	0.619
	LegalΔ	0.904	0.881	0.479	0.924	0.703	0.878	0.856	0.451	0.905	0.674	0.854	0.834	0.427	0.881	0.642
	Issue Tree Rubrics	0.911	0.896	0.488	0.932	0.721	0.885	0.871	0.459	0.914	0.693	0.862	0.848	0.435	0.889	0.658
	JPO	**0.929**	**0.921**	**0.536**	**0.967**	**0.791**	**0.904**	**0.895**	**0.505**	**0.952**	**0.758**	**0.877**	**0.868**	**0.479**	**0.932**	**0.725**
	w/o $R_{\text{structure}}$	0.916	0.909	0.512	0.861	0.755	0.892	0.883	0.484	0.835	0.723	0.865	0.857	0.458	0.812	0.693
	w/o $R_{\text{consistency}}$	0.918	0.908	0.481	0.958	0.682	0.894	0.882	0.453	0.941	0.653	0.868	0.856	0.428	0.921	0.627
	w/o token-level advantage	0.922	0.911	0.519	0.961	0.768	0.898	0.886	0.489	0.945	0.735	0.871	0.859	0.463	0.925	0.704
	w/o adaptive clipping	0.924	0.913	0.522	0.963	0.773	0.899	0.888	0.492	0.947	0.741	0.873	0.861	0.467	0.927	0.708
	w/ uniform token weights	0.919	0.908	0.501	0.956	0.747	0.895	0.883	0.472	0.938	0.715	0.869	0.857	0.448	0.919	0.685
Llama-3.2-3B-Instruct	Pre-trained	0.438	0.365	0.108	0.273	0.184	0.412	0.342	0.096	0.245	0.168	0.389	0.318	0.085	0.221	0.155
	RL-only	0.761	0.722	0.284	0.658	0.458	0.736	0.698	0.264	0.631	0.435	0.713	0.676	0.248	0.602	0.413
	SFT	0.856	0.838	0.384	0.882	0.607	0.831	0.814	0.363	0.861	0.582	0.808	0.792	0.342	0.835	0.554
	Vanilla PPO	0.871	0.854	0.445	0.904	0.652	0.846	0.829	0.421	0.883	0.625	0.822	0.807	0.398	0.858	0.596
	LegalΔ	0.882	0.865	0.459	0.917	0.675	0.857	0.841	0.434	0.896	0.647	0.834	0.819	0.411	0.871	0.618
	Issue Tree Rubrics	0.889	0.877	0.468	0.925	0.693	0.864	0.852	0.443	0.905	0.665	0.841	0.829	0.421	0.879	0.635
	JPO	**0.908**	**0.895**	**0.502**	**0.957**	**0.755**	**0.883**	**0.871**	**0.474**	**0.938**	**0.724**	**0.859**	**0.848**	**0.451**	**0.912**	**0.691**
	w/o $R_{\text{structure}}$	0.893	0.881	0.478	0.844	0.718	0.868	0.857	0.451	0.821	0.687	0.845	0.834	0.428	0.796	0.658
	w/o $R_{\text{consistency}}$	0.895	0.879	0.455	0.941	0.634	0.871	0.855	0.429	0.921	0.605	0.847	0.832	0.407	0.896	0.581
	w/o token-level advantage	0.899	0.886	0.485	0.943	0.729	0.875	0.862	0.457	0.924	0.698	0.851	0.839	0.435	0.898	0.669
	w/o adaptive clipping	0.901	0.888	0.491	0.945	0.734	0.877	0.864	0.463	0.926	0.704	0.853	0.841	0.441	0.901	0.674
	w/ uniform token weights	0.896	0.881	0.478	0.938	0.719	0.872	0.858	0.451	0.919	0.689	0.848	0.835	0.429	0.895	0.661

Table 1: Main results on JPO-Dataset, CAIL2018, and LawBench. We report article F1, charge F1, sentence score, 4-Step Completeness, and Full-Chain Consistency.

Dataset	SFT Train	RL Train	Test
JPO-Dataset	239,515	9,691	20,396
CAIL2018	–	–	30,000
LawBench	–	–	1,500

Table 2: Dataset statistics used in our experiments.

latter is computed by averaging the three local consistency scores introduced in Section 3.6. Formal definitions of the two reasoning-oriented metrics are provided in Appendix H.

Training Details. For structured SFT, we train for 2 epochs with batch size 256. The learning rate is 2×10^{-5} for 3B/4B models and 1×10^{-5} for 7B/8B models. For reinforcement learning, we use learning rate 1×10^{-6}, 4 epochs, train batch size 1024, PPO mini-batch size 256, KL loss weight 10^{-3}, and group size 5. Unless otherwise stated, the reward coefficients are set to $\alpha = 0.75$, $\beta = 0.0625$, and $\gamma = 0.1875$, the token-level advantage scaling factor is $\psi = 0.6$, the entropy–logic mixing coefficient is $\zeta = 0.5$, and the base clipping coefficient is $\epsilon = 0.2$. For all JPO variants, reinforcement learning starts from the corresponding SFT checkpoint.

4.2 Main Results

Table 1 reports the unified results. Overall, JPO consistently improves over both the pre-trained and structured SFT baselines across all open-source backbones and all three datasets. The gains appear not only on article prediction, charge prediction, and sentence prediction, but also on 4-Step Completeness and Full-Chain Consistency, indicating that JPO improves both final judgment quality and intermediate reasoning coherence.

Structured SFT already yields large improvements over the pre-trained models, showing that explicit four-step supervision is useful for criminal judgment prediction. JPO still provides clear gains over SFT across all backbones, suggesting that second-stage policy optimization further improves legal grounding and cross-step consistency. The gains are especially clear on compact and mid-sized open-source models, and are particularly consistent on sentence prediction and Full-Chain Con-

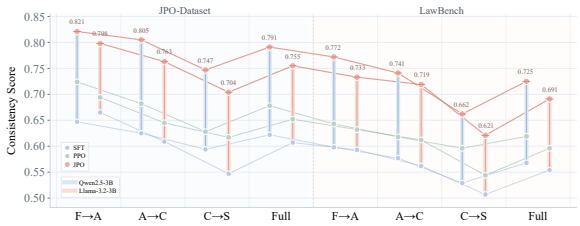

Figure 2: Reasoning-chain breakdown on two representative 3B backbones. F→A, A→C, and C→S denote fact-to-article, article-to-charge, and charge-to-sentence consistency, respectively.

sistency. The proprietary systems remain useful zero-shot references, but are not directly comparable training baselines.

4.3 Detailed Analysis on 3B Backbones

The lower part of Table 1 provides a more detailed comparison on two representative 3B backbones. First, the two-stage design is important: RL-only improves over the pre-trained model, but remains clearly weaker than SFT and JPO, indicating that reinforcement learning without structured initialization is insufficient for stable legal reasoning generation. Second, JPO consistently outperforms strong post-training baselines, including Vanilla PPO, LegalΔ, and Issue Tree Rubrics, showing that the gains do not come from reinforcement learning alone, but from the legal-task-specific reward design and token-aware optimization strategy.

Ablation results further show that each component contributes to the final performance. Removing $R_{\text{structure}}$ most strongly affects 4-Step Completeness, while removing $R_{\text{consistency}}$ causes the largest degradation in sentence score and Full-Chain Consistency. Removing token-level advantage reweighting or adaptive clipping also leads to consistent declines, and replacing importance-aware token weights with uniform token weights produces a larger overall drop. These results support both the reward design and the token-aware optimization strategy.

4.4 Further Analysis

To better localize the gains, Figure 2 decomposes reasoning quality into three local transitions: fact-to-article consistency, article-to-charge consistency, and charge-to-sentence consistency. Across both 3B backbones, JPO improves all three transitions over SFT and Vanilla PPO, with the largest gains appearing in charge-to-sentence consistency. This pattern is consistent with the stronger improvements on sentence prediction and Full-Chain Consistency in Table 1.

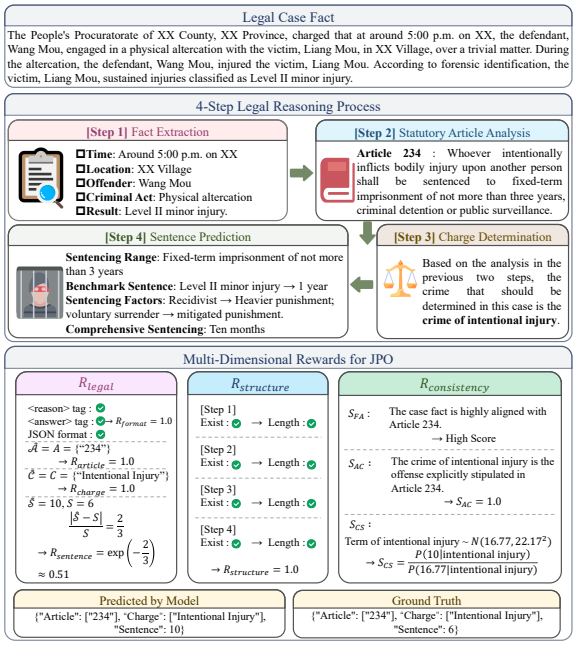

Figure 3: Representative qualitative example illustrating the two-stage process of JPO on a criminal case.

Figure 3 presents a representative qualitative example illustrating the two-stage process of JPO on a criminal case. The figure shows how structured supervised fine-tuning first establishes a four-step reasoning scaffold, and how reinforcement learning further refines the reasoning chain through rewards for legal accuracy, reasoning completeness, and cross-step consistency. Additional qualitative cases and training-dynamics analysis are provided in the appendix.

5 Conclusion

We presented **Juris Policy Optimization (JPO)**, a two-stage post-training framework for structured legal reasoning in Chinese criminal judgment prediction. JPO combines structured supervision with reinforcement learning to optimize legal prediction quality, reasoning completeness, and cross-step consistency, while further improving long-form legal generation through token-level advantage reweighting and adaptive policy clipping. Experiments on multiple open-source backbones and three Chinese legal benchmarks show that JPO consistently improves both final judgment prediction and reasoning-oriented metrics over strong baselines, suggesting that legal post-training should optimize not only final answers, but also the intermediate dependency structure connecting facts, statutes, charges, and sentencing outcomes.

Limitations

JPO is evaluated on Chinese criminal judgment prediction with a structured reasoning chain from facts to statutory articles, charges, and sentencing outcomes. While this setting captures an important class of legal reasoning problems, it does not cover all aspects of judicial decision-making, such as procedural issues or more complex case structures.

In addition, the reasoning-oriented rewards and evaluation metrics in JPO rely on lightweight proxy signals rather than expert-annotated legal reasoning. These signals provide practical and stable supervision for post-training, but they should not be interpreted as complete substitutes for professional legal assessment.

Finally, the structured rationales used in Stage I are generated by a stronger teacher model and may inherit imperfections of model-generated supervision. More broadly, JPO is intended as a research framework for legal NLP, and any practical use in legal settings would still require careful human oversight and broader validation.

Ethics Statement

This work studies structured legal reasoning for Chinese criminal judgment prediction using public judicial documents. Although the data are publicly available, they may still contain sensitive information and historical biases.

The structured rationales and reasoning-oriented rewards used in JPO are intended as weak supervision and computational proxies, rather than authoritative legal explanations or substitutes for expert legal judgment.

JPO is developed for legal NLP research and should not be used as an autonomous decision system in real legal practice. Any practical use would require careful human oversight and broader validation.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Hua Cai, Shuang Zhao, Liang Zhang, Xuli Shen, Qing Xu, Weilin Shen, Zihao Wen, and Tianke Ban. 2025. Unilaw-r1: A large language model for legal reasoning with reinforcement learning and iterative inference. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18128–18142.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Xin Dai, Buqiang Xu, Zhenghao Liu, Yukun Yan, Huiyuan Xie, Xiaoyuan Yi, Shuo Wang, and Ge Yu. 2025. LegalΔ: Enhancing legal reasoning in llms via reinforcement learning with chain-of-thought guided information gain. *arXiv preprint arXiv:2508.12281*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, and 1 others. 2024. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7933–7962.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Jinu Lee, Kyoung-Woon On, Simeng Han, Arman Cohan, and Julia Hockenmaier. 2025. Evaluating legal reasoning traces with legal issue tree rubrics. *arXiv preprint arXiv:2512.01020*.

Pengjie Liu, Zhenghao Liu, Xiaoyuan Yi, Liner Yang, Shuo Wang, Yu Gu, Ge Yu, Xing Xie, and Shuanghua Yang. 2024. Legalduet: Learning effective representations for legal judgment prediction through a dual-view legal clue reasoning. *arXiv preprint arXiv:2401.15371*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin,

Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Weijie Shi, Han Zhu, Jiaming Ji, Mengze Li, Jipeng Zhang, Ruiyuan Zhang, Jia Zhu, Jiajie Xu, Sirui Han, and Yike Guo. 2025. Legalreasoner: Step-wised verification-correction for legal judgment reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7297–7313.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. arxiv 2023. *arXiv preprint arXiv:2302.13971*, 10.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and 1 others. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Additional Experimental Details

This appendix provides additional details for the experimental section, including JPO-Dataset construction, data preprocessing, structured rationale generation, output parsing and evaluation protocol, formal definitions of reasoning-oriented metrics, implementation details of consistency scoring, baseline adaptation, training dynamics, extended ablations, hyper-parameter sensitivity, efficiency analysis, supplementary qualitative examples, and error analysis.

B JPO-Dataset Construction

JPO-Dataset is constructed from criminal judgment documents collected from China Judgments Online. We build this dataset to better reflect more recent criminal case distributions and legal expressions than widely used earlier Chinese criminal judgment prediction benchmarks, which were derived from relatively older public judicial documents.

Following the general construction protocol of CAIL2018, we extract four core fields from each judgment document: fact descriptions, applicable statutory articles, charges, and sentencing outcomes. We then apply preprocessing and normalization to obtain a unified prediction schema across cases.

Document collection and filtering. We collect public criminal cases and retain judgments that contain sufficiently complete factual descriptions and explicit judicial outcomes. We exclude documents with missing key fields, duplicated records, or severely truncated facts. To keep the setting focused and comparable to prior criminal judgment prediction benchmarks, we restrict the current study to single-defendant cases with a normalized fact-to-judgment structure.

Field extraction and normalization. For each case, we extract the fact section as model input and use the judicial outcome section to derive the target labels. We normalize article references to unified statutory identifiers, map charge expressions to canonical charge labels, and convert sentencing expressions into normalized sentence values. This step reduces surface-form variation across documents and makes the prediction targets consistent across cases.

Sentence normalization. Sentencing outcomes are converted into a normalized scalar value used for both training and evaluation. This normalization mainly resolves surface-form variation in judicial documents, such as different expressions of imprisonment length and equivalent textual formulations. The resulting value is used as the target S in the main paper.

Split usage. The resulting dataset is used for both structured supervised fine-tuning and reinforcement learning. The SFT split is used to learn the four-step reasoning template together with the final prediction format, while the RL split is used for post-training under the composite reward. Table 2 in the main paper reports the final split sizes.

Scope and limitations. JPO-Dataset is designed for criminal judgment prediction under a structured reasoning formulation. It does not cover all judicial scenarios, such as multi-defendant cases, highly atypical fact patterns, or complex procedural decisions. The dataset therefore serves as a focused benchmark for structured criminal judgment prediction rather than a complete representation of judicial decision-making in practice.

C Dataset Profile

To provide a clearer picture of the evaluation setting, we summarize several basic statistics of JPO-Dataset beyond the split sizes reported in the main paper. These statistics are intended to describe the scale and diversity of the dataset, including the temporal scope of the collected cases, the coverage of charge and article labels, the length of fact descriptions, and the distribution of normalized sentencing outcomes.

JPO-Dataset is constructed from public criminal judgment documents collected over a recent time range. After filtering and normalization, the dataset covers a diverse set of criminal charges and statutory articles, while remaining focused on single-defendant criminal cases. Compared with earlier benchmark-style datasets, its main purpose is not to maximize label coverage, but to provide a cleaner and more recent testbed for structured criminal judgment prediction.

These statistics help contextualize the task difficulty. In particular, the number of distinct charge and article labels reflects the breadth of the prediction space, while fact length indicates the complexity of the input reasoning context. Sentence statistics provide a coarse description of the target distribution used for the normalized sentencing prediction setting.

Statistic	Value
Time range of collected cases	2024–2026
Number of unique charges	192
Number of unique statutory articles	176
Average fact length (tokens)	217.1
Median fact length (tokens)	194
Average number of articles per case	1.04
Average number of charges per case	1.02
Average normalized sentence value	14.0
Median normalized sentence value	7

Table 3: Profile statistics of JPO-Dataset. Replace the placeholder values with the final dataset statistics.

D Structured Rationale Generation for SFT

Most legal judgment prediction datasets do not provide explicit intermediate reasoning supervision. To construct structured SFT targets, we use a stronger instruction-following teacher model to generate four-step legal rationales aligned with our reasoning template.

Rationale structure. For each case, the teacher is instructed to generate four ordered segments: (1) fact extraction, (2) statutory analysis, (3) charge determination, and (4) sentence prediction. Each segment is required to be concise, legally relevant, and aligned with the final normalized answer block.

Target construction. The student target consists of the teacher-generated four-step reasoning trace together with the gold judicial outcome. This design allows the model to learn not only the output format, but also a stable intermediate reasoning scaffold before reinforcement learning.

Role of teacher rationales. The teacher rationales are used as weak structured supervision rather than as gold legal explanations. Their purpose is to expose the student model to a consistent and interpretable reasoning template. The second-stage reinforcement learning objective then further refines whether the generated reasoning remains correct, complete, and cross-step coherent.

E Output Format and Parsing Protocol

Because JPO jointly evaluates final outcomes and intermediate reasoning, we use a unified output schema during both training and evaluation.

Structured response format. Each response contains two parts: a four-step reasoning trace and

a final prediction block. The reasoning trace is organized into the four predefined stages, while the final prediction block contains normalized article, charge, and sentence predictions.

Parsing of final answers. Predicted statutory articles and charges are extracted from the final prediction block and mapped to the normalized label space used by the dataset. Sentence predictions are parsed into the normalized scalar value used for evaluation. Outputs that fail to match the required schema receive a lower format reward and may also produce zero scores for the corresponding set-based or consistency-based terms when parsing fails.

Empty-set handling. For reward computation, when the predicted article set $\hat{A}$ is empty, we define the corresponding set-based quantities to be zero. Similarly, when the predicted charge set $\hat{C}$ is empty, the corresponding charge-related reward and consistency terms are set to zero. This avoids undefined behavior in set-based rewards and consistency scores.

F Implementation Details of Consistency Scoring

The main paper defines three local consistency scores: fact-to-article consistency S_{FA}, article-to-charge consistency S_{AC}, and charge-to-sentence consistency S_{CS}. Here we provide additional implementation details.

Fact-to-article consistency. The Naive Bayes article predictor used in S_{FA} is trained on the training corpus using fact-side lexical features and article labels. Its role is not to serve as a standalone predictor, but to provide a lightweight estimate of whether the predicted articles are supported by the observed facts.

Article-to-charge consistency. The article–charge association matrix used in S_{AC} is built from rule-based associations derived from statutory references and charge annotations in the training data. It is intended as a lightweight compatibility table rather than a complete codification of legal doctrine.

Charge-to-sentence consistency. For S_{CS}, we estimate a charge-conditioned sentence distribution from the training data. The purpose of this term is to measure whether the predicted sentence lies in a plausible region conditioned on the predicted charge. It should therefore be interpreted as a statistical plausibility signal rather than a normative sentencing model.

Interpretive boundary. All three consistency scores are designed as computable proxies that support policy optimization and reasoning-oriented evaluation. They are not intended to fully represent expert legal analysis or replace professional legal judgment.

G Baseline Adaptation

To improve experimental fairness, all trainable open-source baselines are adapted to the same underlying prediction setting.

Shared task interface. All trainable methods are evaluated under the same structured output schema and the same normalized target space for articles, charges, and sentences.

Initialization and data usage. For methods involving post-training, reinforcement learning starts from the corresponding SFT checkpoint unless otherwise noted. This keeps the comparison focused on the effect of the post-training objective rather than differences in initialization.

Interpretation of proprietary references. The proprietary systems reported in the main paper are included only as zero-shot references. Since they are not post-trained on JPO-Dataset, they should not be interpreted as directly comparable training baselines.

H Definitions of Reasoning-Oriented Metrics

We evaluate reasoning quality with two process-oriented metrics beyond final label accuracy.

4-Step Completeness. A generated response is counted as complete if it explicitly contains all four reasoning stages required by our framework: fact extraction, statutory analysis, charge determination, and sentence prediction. In practice, we detect these stages using the structured output template and require each segment to be non-empty. The final 4-Step score is the average completeness rate over the evaluation set.

Full-Chain Consistency. Full-Chain Consistency measures the overall coherence of the generated legal reasoning chain. For each response, we first compute three local consistency scores

introduced in Section 3.6: fact-to-article consistency S_{FA}, article-to-charge consistency S_{AC}, and charge-to-sentence consistency S_{CS}. We then average these three scores:

$$\text{FullChain} = \frac{S_{FA} + S_{AC} + S_{CS}}{3}. \tag{30}$$

The final Full-Chain score is obtained by averaging this value over all evaluation samples.

Interpretation. Article F1, charge F1, and sentence score evaluate the final judgment outcome. By contrast, 4-Step Completeness measures whether the model follows the expected reasoning structure, while Full-Chain Consistency measures whether the reasoning process remains coherent across stages according to the three local consistency proxies defined in the main paper. These metrics are intended to capture useful process signals for structured legal reasoning, rather than to serve as complete substitutes for expert legal evaluation.

I Training Dynamics

Figure 4 illustrates how the gains of JPO emerge during training on Qwen2.5-3B. During the SFT stage, the most immediate improvements appear in output validity and coarse legal prediction, reflecting that structured supervision mainly teaches the model to follow the desired reasoning template. During the RL stage, the most visible gains shift to sentence prediction and Full-Chain Consistency, indicating that reinforcement learning primarily improves legal coherence and decision calibration beyond template acquisition alone.

This pattern is consistent with the design of JPO. The first stage mainly stabilizes the reasoning format and teaches the model to expose intermediate legal steps, while the second stage focuses on refining whether these steps remain mutually compatible. In particular, the later gains in sentence prediction suggest that policy optimization is especially useful for improving downstream reasoning quality after the basic structure has already been acquired.

The main paper integrates the core ablation results into the unified main table. Here we further provide a finer-grained ablation analysis for the consistency components on Qwen2.5-3B. The results are shown in Table 4. Among the three local consistency transitions, removing charge-to-sentence consistency causes the largest drop in

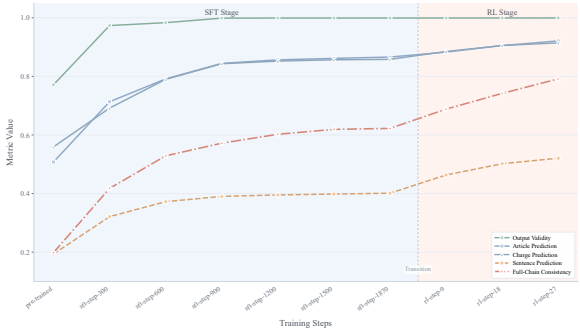

Figure 4: Training dynamics of JPO on the representative Qwen2.5-3B backbone. The SFT stage mainly improves output validity and coarse prediction quality, while the RL stage yields the largest gains in sentence prediction and reasoning consistency.

Method	Art. F1	Charge F1	Sent.	Full-Chain
JPO	0.929	0.921	0.536	0.791
w/o S_{FA}	0.923	0.915	0.520	0.768
w/o S_{AC}	0.920	0.908	0.513	0.755
w/o S_{CS}	0.925	0.916	0.497	0.763
Entropy-only weighting	0.923	0.912	0.523	0.769
Logic-only weighting	0.924	0.914	0.525	0.771
w/o KL regularization	0.925	0.915	0.515	0.741

Table 4: Extended ablation analysis on Qwen2.5-3B. Green highlights the full JPO model, and light beige highlights key ablations.

sentence prediction, while removing fact-to-article consistency most directly harms statutory grounding.

These results further support the design intuition of JPO. The three consistency components do not contribute equally: earlier transitions mainly affect whether the model selects legally plausible statutory support, whereas the later charge-to-sentence component plays a more direct role in sentence calibration. The weighting variants also show that combining entropy-based uncertainty with legal-logic-aware importance is more effective than using either signal alone.

J Extended Hyper-Parameter Sensitivity

We evaluate the robustness of JPO with respect to several key hyper-parameters on Qwen2.5-3B using JPO-Dataset.

J.1 Reward Weight Sensitivity

Table 5 shows the effect of varying the reward composition. The best overall trade-off is obtained when legal prediction remains the dominant term while structure and consistency provide auxiliary guidance. Over-emphasizing any single component leads to weaker overall performance, suggesting that JPO benefits from balancing final-answer

α	β	γ	Art. F1	Charge F1	Sent.
0.85	0.05	0.10	0.923	0.913	0.518
0.75	0.0625	0.1875	0.929	0.921	0.536
0.65	0.0625	0.2875	0.926	0.917	0.530
0.33	0.33	0.33	0.904	0.893	0.494

Table 5: Sensitivity to reward weight composition. The best setting is lightly highlighted.

ψ	Art. F1	Charge F1	Sent.
0.0	0.922	0.911	0.519
0.3	0.924	0.915	0.523
0.6	0.929	0.921	0.536
0.9	0.927	0.918	0.531

Table 6: Sensitivity to token-level advantage scaling ψ.

supervision and process-level reasoning supervision.

J.2 Token-Level Advantage Scaling

Table 6 shows the effect of the token-level advantage scaling factor ψ. Moderate reweighting performs best, while either removing token-level scaling or applying excessively strong scaling leads to weaker results. This supports the use of importance-aware but still conservative token-level optimization.

J.3 Entropy–Logic Mixing

Table 7 shows the effect of the entropy–logic mixing coefficient ζ. The best setting lies near the middle of the range, indicating that both uncertainty information and legal-logic-aware stage information contribute to token importance estimation. Using only one of the two signals leads to weaker performance.

K Training Efficiency and Stability

Table 8 reports the efficiency and stability comparison on Qwen2.5-3B. JPO introduces moderate additional training cost relative to SFT, but yields substantially stronger reasoning quality and lower variation than reward-only PPO.

The efficiency results suggest that the additional complexity of JPO is practically manageable: compared with vanilla PPO, the increase in GPU hours is limited, while the gains in reasoning quality and stability are substantial. This trade-off is favorable for legal-domain post-training, where stable structured generation is often more important than marginal reductions in training cost.

L Additional Qualitative Examples

In the main paper, we use a compact flow-style figure to visualize a representative case. Here we

ζ	Art. F1	Charge F1	Sent.
0.0	0.924	0.914	0.525
0.25	0.926	0.917	0.529
0.5	0.929	0.921	0.536
0.75	0.925	0.916	0.527
1.0	0.923	0.912	0.523

Table 7: Sensitivity to entropy–logic mixing coefficient ζ.

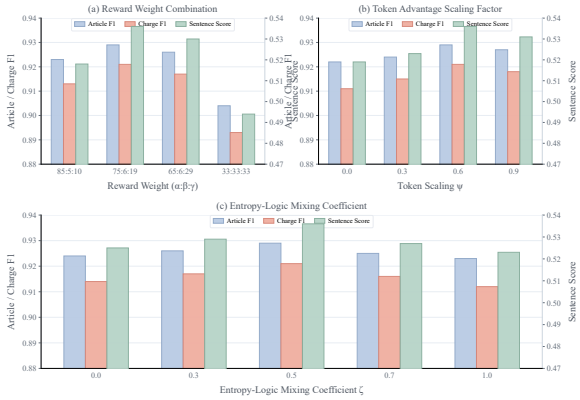

Figure 5: Hyper-parameter sensitivity of JPO on the representative Qwen2.5-3B backbone.

provide additional qualitative observations based on detailed generated outputs.

Example A: Better statutory grounding. In repeated narcotics-trafficking cases, the SFT model often predicts the correct charge but gives only shallow statutory justification. JPO more reliably connects trafficking frequency, drug quantity, and statutory threshold to the cited articles, yielding a better-aligned fact-to-article transition.

Example B: Improved sentence calibration. In theft and fraud cases with confession but limited restitution, SFT tends to under-calibrate the sentence. JPO more often predicts a sentence that falls in a plausible charge-conditioned region and better explains why mitigation is limited, producing a more coherent charge-to-sentence transition.

Example C: Stronger multi-stage coherence. In cases with highly similar facts but different legal thresholds, JPO is more likely than SFT to preserve coherence from fact extraction to statute selection and then to charge determination, instead of jumping directly to the final answer.

Interpretive caution. These qualitative examples are intended to illustrate the types of improvements encouraged by JPO, rather than to serve as formal legal evaluations. They should therefore be interpreted as complementary evidence alongside the quantitative metrics reported in the main paper.

Method	GPU Hours	Avg. Length	Std. Full-Chain
SFT	34	456	0.018
Vanilla PPO	49	471	0.025
JPO	53	478	0.015

Table 8: Training efficiency and stability on Qwen2.5-3B.

Error Type	Proportion
Missing secondary article	27.5%
Correct charge but wrong sentence band	25.3%
Confusion between neighboring charges	21.6%
Incomplete mitigation/aggravation reasoning	16.7%
Formatting or parsing errors	8.9%

Table 9: Main remaining error types of JPO.

M Error Analysis

Table 9 summarizes the main remaining error types of JPO. The most common failures involve secondary article omission, difficult sentence calibration under fine-grained aggravating and mitigating factors, and confusion between neighboring charges with similar fact patterns.

The error distribution suggests that the main remaining challenge is no longer gross formatting failure, but fine-grained legal distinction. In particular, secondary article omission and difficult sentence calibration indicate that even when the model captures the broad reasoning chain correctly, it can still miss more subtle statutory interactions or sentencing adjustments. This observation is consistent with the fact that later-stage legal reasoning remains the hardest part of the task.

N Prompt Template for Structured SFT

Figure 6 shows the prompt template used to elicit four-stage legal rationales from the teacher model during structured SFT.

Prompt Used for Teacher Rationale Generation

Given a legal case information and the corresponding judgment result, generate a reasoning process explaining how to deduce the judgment result from the legal case information. The reasoning process you generate must strictly follow the following four steps:

Step 1 Fact Extraction: Review the legal case information and extract key factual elements.

Step 2 Statutory Analysis: Based on the key factual elements extracted in Step 1, identify the relevant provisions of the Criminal Law, and explain the correspondence between the elements of these provisions and the facts of the case.

Step 3 Charge Determination: Integrate the analysis from Step 1 and Step 2 to determine the charge that should be identified in this case.

Step 4 Sentence Prediction: First, based on the charge identified in Step 3 and the legal provisions determined in Step 2, specify the applicable statutory sentencing range for this case. Then, based on the criminal facts, preliminarily determine the benchmark sentence within the statutory sentencing range. Finally, consider the sentencing circumstances in the case to comprehensively provide the predicted sentence and explain the reasons..

Output: The student model is trained to generate both the structured reasoning trace and the final normalized prediction block.

Figure 6: Prompt template used to generate four-stage legal rationales for structured supervised fine-tuning.

Case 1:

Fact: The prosecution alleged that on a certain afternoon in XX month XXXX year, the defendant Li Mouyi and his wife Ni Mou, at their residence near No. XX, XX Community, XXX Village, XX Street, XX City, had a dispute with their neighbor Li Moujia and his wife Du Mou over the placement of drainage pipes for a village road construction project. During the altercation, the defendant Li Mouyi pushed the victim Li Moujia, causing him to fall and fracture his right leg. Forensic examination determined that the victim Li Moujia suffered minor injuries, with a post-injury sequela assessed as a Level 10 disability.
Ground Truth: {"Article": ["234"], "Charge": ["Intentional Injury"], "Sentence": 12}

[Step 1]
- **Time:** XX month XXXX year, afternoon
- **Location:** Near No. XX, XX Community, XXX Village, XX Street, XX City
- **Parties:** Defendant Li Mouyi and his wife Ni Mou; Victim Li Moujia and his wife Du Mou
- **Cause:** Dispute over the placement of drainage pipes for a village road construction project
- **Event:** Li Mouyi argued with Li Moujia and Du Mou, then pushed Li Moujia, causing him to fracture his right leg
- **Injury:** Forensic examination: minor injuries, Level 10 disability

[Step 2]
- **Legal Provision:** Article 234 of the Criminal Law of the People's Republic of China
- **Elements:**
  - **Intentional injury to another's body:** In this case, Li Mouyi intentionally pushed Li Moujia, causing him to fracture his right leg, which meets the element of "intentional injury to another's body."
  - **Causing minor injury:** Forensic examination determined that Li Moujia suffered minor injuries, meeting the element of "causing minor injury."

[Step 3]
- **Charge:** Intentional injury

[Step 4]
- **Statutory sentencing range:** According to Article 234 of the Criminal Law, intentional injury causing minor injury is punishable by up to three years imprisonment, criminal detention, or public surveillance.
- **Baseline sentence:** In this case, Li Moujia suffered minor injuries with Level 10 disability, which is considered a moderate degree of minor injury. The initial baseline sentence is estimated at around one year.
- **Sentencing circumstances:** The defendant Li Mouyi voluntarily surrendered and truthfully confessed his crime after returning to the case, constituting voluntary surrender. According to Article 67 of the Criminal Law, voluntary surrender may lead to a lighter or mitigated punishment.
- **Predicted sentence:** Considering the above factors, Li Mouyi's statutory sentence is up to three years. With the voluntary surrender, a lighter punishment is possible. The final predicted sentence is 10 months.

[Answer] {"Article": ["234"], "Charge": ["Intentional Injury"], "Sentence": 10}

Metadata: $Score_article = 1$, $Score_charge = 1$, $Score_sentence = \exp(-3 \times 1/6) \approx 0.61$

Case 2:

Fact: It was established through investigation that at a certain hour on XX day of XX month XXXX year, the defendant Liu went to cut grass in the hills surrounding XXX Village. Upon seeing the victim Mo cutting down two small trees for firewood, which were tea seedlings from his responsibility mountain forest, he argued with the victim. Liu then used the back of his sickle blade to strike Mo's left forehead. Mo suffered a comminuted skull fracture and cerebral contusion (Grade 1 minor injury), an orbital fracture (Grade 2 minor injury), and a laceration on the left eyebrow (minor injury). Liu's family reached a criminal settlement agreement with Mo. Liu compensated Mo for medical expenses, lost wages, nursing fees and disability compensation. Mo expressed forgiveness for Liu's actions.
Ground Truth: {"Article": ["234"], "Charge": ["Intentional Injury"], "Sentence": 5}

[Step 1]
- **Time and Location:** XX hour on XX day of XX month XXXX year, surrounding hills of XXX Village
- **Perpetrator:** Defendant Liu
- **Victim:** Victim Mo
- **Specific act:** Liu used the back of his sickle blade and strike Mo's left forehead.
- **Injury result:** Forensic examination: comminuted skull fracture (Grade 1 minor injury), cerebral contusion (Grade 1 minor injury), orbital fracture (Grade 2 minor injury), laceration on the left eyebrow (minor injury)
- **Settlement:** Liu's family reached a criminal settlement agreement with Mo, compensating 22,000 yuan, and Mo expressed forgiveness

[Step 2]
- **Legal Provision:** Article 234 of the Criminal Law
  - **Paragraph 1:** Intentional injury to another's body is punishable by up to three years imprisonment, criminal detention, or public surveillance.
  - **Paragraph 2:** Whoever commits the crime in the preceding paragraph and causes serious injury shall be sentenced to fixed-term imprisonment of not less than three years but not more than ten years.

[Step 3]
- **Charge:** According to Article 234 of the Criminal Law, Liu Moumou's actions constitute the crime of intentional injury.

[Step 4]
1. **Statutory sentencing range:** According to Article 234, intentional injury to another's body is punishable by up to three years imprisonment.
2. **Baseline sentence:** In this case, Mo's injuries were Grade 1 and Grade 2 minor injuries, within the range of up to three years imprisonment.
3. **Sentencing circumstances:**
   - **Mitigating factors:** Liu's family reached a criminal settlement agreement with Mo, compensating 22,000 yuan, and Mo expressed forgiveness.
   - **Aggravating factors:** No significant aggravating factors.
4. **Comprehensive assessment:** Considering Liu's actions caused Grade 1 and Grade 2 minor injuries to Mo, and his active compensation and receipt of forgiveness after the incident, a lighter punishment within the statutory range is possible.
5. **Predicted sentence:** 8 months

[Answer] {"Article": ["234"], "Charge": ["Intentional Injury"], "Sentence": 8}

Metadata: $Score_article = 1$, $Score_charge = 1$, $Score_sentence = \exp(-3 \times 3/5) \approx 0.17$